Endogenous Spatial Regression and Delineation of Submarkets

Endogenous spatial regression and delineation of submarkets:

A new framework with application to housing markets [#]

Arnab Bhattacharjee,[*] Heriot-Watt University, UK. a.bhattacharjee@hw.ac.uk

Eduardo Castro, University of Aveiro, Portugal. ecastro@ua.pt

Taps Maiti, Michigan State University, USA. maiti@stt.msu.edu

João Marques, University of Aveiro, Portugal. jjmarques@ua.pt

Abstract

Housing submarkets have been defined by different criteria: i) similarity in house attributes; ii) similarity in hedonic prices; or iii) substitutability of houses. We show that spatial clustering on i) and ii) also satisfies criterion iii), and develop inferences based on functional linear regression of a hedonic house price model. Then, we delineate submarkets by clustering (jointly) on the surfaces of the estimated functional partial effects and housing features. The above model incorporates both spatial heterogeneity and endogenous spatial dependence. Application to an urban conglomeration in Portugal implies submarkets that emphasize the historical and endogenous evolution of urban spatial structure.

KEYWORDS: Spatial heterogeneity; Endogenous spatial dependence; Housing submarkets; Spatial lag model; Geographically weighted regression; Functional linear regression.

JEL CLASSIFICATION: C21; R31; C38; C51.

[#] We thank the Editor and two anonymous reviewers for their encouraging, constructive and challenging comments that helped us improve the paper substantially. The paper has also benefited from comments by participants at the Workshop on Housing Economics (European Network for Housing Research, Vienna, March 2012) and the VI World Conference of the Spatial Econometrics Association (Salvador, Brazil, July 2012) , and seminars at London School of Economics and University of Illinois at Urbana-Champaign. We are grateful to Anil Bera, Atanas Christev, Edwin Deutsch, Duncan Maclennan, Daniel McMillen, Peter Robinson, Mark Schaffer and Jacques-Francois Thisse for many helpful comments and suggestions. The authors acknowledge financial support from the Portuguese Foundation for Science and Technology (FCT) on the project DONUTS (PTDC/AURURB/100592/2008), together with the Competitiveness Factors Thematic Operational Programme (COMPETE) of the Community Support Framework III (European Commission) and the European Community Fund FEDER. The real estate portal Casa Sapo is thanked for permission to use their data for the empirical analyses. The usual disclaimer applies.

[*] Correspondence: Arnab Bhattacharjee, Heriot-Watt University, Spatial Economics and Econometrics Centre (SEEC), Room 1.06, Mary Burton Building, Edinburgh EH14 4AS, Scotland, United Kingdom. E-mail: a.bhattacharjee@hw.ac.uk. Tel.: +44 (0)131 4513482. Fax: +44 (0)131 4513296.

"(Social) space is a (social) product ... the space thus produced also serves as a tool of thought and of action; that in addition to being a means of production it is also a means of control, and hence of domination, of power. ... Change life! Change Society! These ideas lose completely their meaning without producing an appropriate space." (Lefebvre, 1974 [1991], p.26, p.59).

1. Introduction

Definition of housing submarkets is important at both conceptual and empirical levels. Endogenous evolution of space, as emphasized by Lefebvre (1974 [1991]), is central in this context, and more generally to spatial dynamics in urban housing markets. Understanding endogenous housing segmentation enables researchers to study spatial variation in housing prices, improving lenders' and investors' abilities to price the risk associated with financing homeownership; at the same time it reduces search costs to housing consumers (Malpezzi, 2003, Goodman and Thibodeau, 2007). By its very nature, housing is a heterogeneous good, characterized by a diverse set of attributes (Lancaster, 1966; Rosen, 1974) and segmented and structured by complex spatial patterns. Different social groups, with specific tastes, preferences and economic capabilities tend to be organized into distinct territorial clusters (Galster, 2001). However the literature does not suggest an unequivocal and unique spatial approach to analyse this issue, encompassing different philosophies, techniques and criteria.

Housing markets are complex. Rather than being defined by a single combination of a quantity and a price, the market equilibrium for a heterogeneous good such as a house is given by the combination of a vector of hedonic characteristics with a vector of hedonic prices (Lancaster, 1966; Rothenberg *et al.*, 1991). A unique vector of hedonic prices, combined with a distribution of houses with different hedonic characteristics, is a necessary condition for the existence of a single equilibrium and a unique market. However, what we generally observe is the co-existence of several submarkets, each corresponding to a different market equilibrium (Rothenberg *et al.*, 1991). This heterogeneity, driven by supply rigidities and transaction costs, shapes the territory as landscapes of submarkets. Such landscapes can be either represented as sets of hedonic functions, each with one particular vector of hedonic prices, or as a continuum of vectors represented by a hedonic functional. This paper focuses on the application of a functional representation to the empirical study of housing hedonic price models.

Because of such inherent heterogeneity over space, understanding housing markets and the conduct of housing policy crucially depends on delineation of submarkets (Rothenberg *et al.*, 1991). Each submarket is characterized by different supply and demand curves and a different equilibrium. A multitude of criteria have

been proposed in the literature for defining housing markets and their constituent submarkets, each based on different theoretical assumptions underlying its definition. There are three main criteria for the definition of submarkets: i) similarity in hedonic characteristics; ii) similarity in hedonic prices; or iii) close substitutability of housing units. We argue that spatial clustering based simultaneously on criteria i) and ii) is a sufficient condition for criterion iii) to hold. Since criterion i) is directly observable, we focus on ii). Thus, the central object of our inference is a regression model where the dependent variable is logarithm of house prices per square meter and housing features are regressors. The partial effect of these housing features varies over a two-dimensional territory. In this paper, we focus on a single regressor, logarithm of living area, so that the functional regression coefficient $\beta(s)$ can be interpreted as an elasticity which reflects a positive but decreasing marginal utility of living area. Generally, $-1 < \beta(s) < 0$; when the elasticity approaches zero consumers show a very low satiation of living space, while a value close to negative unity (-1) reflects a submarket with a rigid demand for living space.

Appropriate characterization of spatial structure is a key element of such analyses. Specifically, three distinct aspects of space – spatial heterogeneity, spatial dependence and spatial scale – are central to understanding the spatial organization of housing submarkets. Spatial heterogeneity relates to contextual variation over space (Anselin, 1988). In this paper, we consider a spatial cross section context, where spatial heterogeneity is modelled as variation across submarkets in (heterogeneous) slopes and intercepts (spatial fixed effects) of a regression model. By contrast, spatial dependence is associated with spatial spillover, contagion and diffusion, which results in spatial autocorrelation between different units (Anselin, 1988). Additionally, choice of an appropriate spatial scale is important (Malpezzi, 2003), where the choice may range from national or regional scale, through metropolitan areas, to below the metropolitan level. Appropriate modelling of spatial heterogeneity depends on the choice of scale: the correct choice increases prediction accuracy of the estimated hedonic models and, in many cases, negates strong spatial dependence (Pesaran, 2006; Pesaran and Tosetti, 2011).

We show how estimates of a hedonic regression model can be used to identify submarkets, by clustering jointly on the surface of the heterogenous slope $\beta(s)$ and the hedonic features $x(s)$. For this purpose, we propose a new framework to analyse housing markets, based on a synthesis of spatial econometrics, functional data analysis (FDA) and locally (geographically) weighted regression (GWR). We consider a spatial lag model, regressing logarithm of price per square meter of living space on logarithm of house area, allowing for spatial heterogeneity (spatial fixed effects and slope heterogeneity) and endogenous spatial dependence captured by a spatial weights

matrix *W*. This, in turn, leads to a functional regression model where the response variable is scalar and the functional regressor is a spatially weighted version of the average functional surface of the regressor. When kernel weights are used, the model is very similar to GWR. This synthesis of GWR and FDA offers a spatial statistical model that is very rich and enable the full range of spatial analyses of housing markets. This model addresses two main limitations of previous approaches. First, the framework does not require housing submarkets to be fixed *a priori*. They can be delineated *ex post* by spatial (or even non-spatial) clustering of an estimated functional regression slope and hedonic features. Second, estimation and endogeneity of spatial weights can be addressed with the proposed model. Application to the housing market of the Aveiro-Ílhavo urban area in Portugal implies submarkets that emphasize the historical and endogenous evolution of urban space.

The paper is organised as follows. Section 2 discusses the recent spatial econometrics literature applied to the hedonic pricing model, followed by delineation of submarkets in section 3. Section 4 highlights limitations of the spatial econometrics framework, discusses alternative approaches and proposes a new synthesis of several methods. Based on this synthesis, section 5 develops methodology for submarket delineation, followed by an application to the urban housing market of Aveiro and Ílhavo in Section 6. Finally, section 7 concludes. An expanded version of the paper containing further details is included as online supplementary material.

2. Spatial Econometric Hedonic House Price Models

This paper uses hedonic models to study spatial dynamics and house prices. Typically, hedonic and repeated sales models of house prices reflect two spatial features – geographically varying price elasticities and substantial spatial clustering – that typically arise from supply rigidities, search costs and social segregation (Malpezzi, 2003). Such spatial clustering has been explained by neighbourhood characteristics such as crime rates, schooling, transport infrastructure and quality of public services, and social interaction and segregation; see, for example, Rothenberg *et al.* (1991). Therefore, empirical estimation of hedonic housing price models and the use of such estimates for evidence and policy have to take spatial effects explicitly into account.

2.1. Hedonic pricing model

Hedonic pricing models (Lancaster, 1966; Rosen, 1974) are frequently used in housing studies, particularly for valuation of housing attributes, neighbourhood features and access to central and local services, and for construction of price indices based on single sales data; see Malpezzi (2003) for an excellent review. In hedonic pricing models, dwelling unit values (or prices or rents) are regressed on a bundle of characteristics of the house:

$$Y = f(S, N, L, C, T), \quad (1)$$

where *Y* denotes the value of the house (typically logarithm of price, or logarithm of price per unit area), and *S*, *N*, *L*, *C* and *T* denote respectively: **S**tructural characteristics of the dwelling (living space, type of construction, tenure, etc.); **N**eighbourhood characteristics and local amenities; **L**ocation within the market (or access to employment/ business centre); other **C**haracteristics (access to utilities and public services, such as water supply, electricity, central heating, etc.); and the **T**ime when the value is observed. Estimation of a hedonic price function yields implicit prices for housing characteristics that can be interpreted as willingness-to-pay estimates. This allows analysis of various upgrading and policy scenarios, targeted on specific subgroups, defined either by socio-economic characteristics or by location. Thus, the model facilitates understanding of residential location, and therefore urban structure, and provides valuable input towards urban planning and housing policy.

The two main limitations of traditional hedonic models are: (a) the frequent assumption that hedonic prices do not vary spatially; and (b) inadequate attention to spatial spillover effects. To overcome these problems, we consider a hedonic model incorporating both spatial dependence and spatial variation in the relationship between house prices and living space. Following Bhattacharjee *et al.* (2012), we adopt a log-log form, where logarithm of price per square meter of living space is regressed on logarithm of house area, conditioning on several other hedonic housing characteristics, used as control variables and modelled by statistical factor analysis.

2.2. Spatial issues in hedonic pricing estimates

The recent literature has discussed potential bias and loss of efficiency that can result when spatial effects are ignored in the estimation of hedonic models; see, for example, LeSage and Pace (2009), Anselin and Lozano-Gracia (2008) and Anselin *et al.* (2010). Specifically, these biases can result both from inappropriate modelling of endogenous spatial effects and inadequate attention to spatial heterogeneity, while heteroscedasticity and spillovers in unobservable errors lead to inefficiency. Adequate modelling of spatial heterogeneity and spatial dependence is therefore crucial (Anselin, 1988), as well as the choice of an appropriate spatial scale (Malpezzi, 2003). We now turn to a discussion of these spatial issues in the construction of hedonic pricing models.

2.2.1. Spatial scale and housing submarkets

The definition of the most appropriate scale in the analysis of urban spatial patterns is a crucial aspect. The spatial configuration usually varies with scale. A specific urban pattern that is structured at one scale may appear

to be disordered at other scales, leading to the so called "ecological fallacy"; one explanation is the different effects of agglomeration economies that emerge at specific scales (Anas *et al.*, 1998; Fujita and Thisse, 2002).

In practise, the metropolitan area is a common choice because it is usually thought of as a labour market, which in theory is approximately coincident with housing markets (Malpezzi, 2003). However, submarkets below the metropolitan level can be segmented by location (central city/suburb), or by housing quality, or even by race or income levels. Such segmentation informs both the study of residential neighbourhood choice and design of urban housing policy. The urban area of Aveiro has adequate size and variability to study spatial heterogeneity in the shadow (hedonic) price of housing space, and at the same time allow for spatial spillovers in house prices.

2.2.2. Spatial heterogeneity and submarkets

The model for spatial heterogeneity must, in principle, be based on a theoretical framework explaining why and how housing markets are segmented. As discussed above, the literature has defined submarkets either by similarity in hedonic housing characteristics (Rothenberg *et al.*, 1991; Adair *et al.*, 1996; Bourassa *et al.*, 1999; Watkins, 2001), similarity in hedonic prices (Dale-Johnson, 1982; Rothenberg *et al.*, 1991), or close substitutability of housing units (Grigsby *et al.*, 1987; Goodman and Thibodeau 2007; Pryce, 2013).

In the first approach, a submarket is a collection of regions, or housing units located therein, which have similar bundle quality, or supply a similar set of hedonic characteristics. The degree of similarity required is a matter of judgment, particularly since a perfectly homogeneous location may be very small and therefore not useful for estimating hedonic models (Bourassa *et al.*, 2003). In any case, the delineation of submarkets implied by this approach can be directly applied to the data by clustering on hedonic characteristics. In essence, this approach has a logic that stresses the role of branding and social segregation as the driver of submarkets.

The second approach defines submarkets as locations where hedonic (shadow) prices for different features are homogeneous. Submarkets can then be interpreted as clusters of houses with characteristics adjusted to a particular demand behaviour reflected in a set of equilibrium prices. This approach, proposed by Bourassa *et al.* (2003), is intimately related to the basic philosophy of hedonic models stating that, within the same submarket, the implicit prices for each housing feature must be homogeneous. This criterion also improves price predictions.

The third criterion is the degree of substitutability (Grigsby *et al.*, 1987). Pryce (2013) measures substitutability by cross-price elasticities of price at different locations, estimated using a spatial panel regression model where the logarithm of house prices at one location is regressed on log-price at the same location at another time point

together with a time trend.[1]

By contrast with Pryce (2013), this paper takes the view that, in the context of a hedonic house price model based on cross-section data, a collection of suitably chosen housing characteristics constitutes a more natural set of latent factors. By the logic of a hedonic pricing model, one can expect that these factors account for any strong spatial dependence, which would then render the model as containing only weak spatial dependence.

The above approaches are not necessarily compatible. One can envisage situations where homogeneity in hedonic characteristics does not imply close substitutability. If two locations with similar houses, similar provision of local services and amenities and similar accessibility to the centre are inhabited by two different social groups, it is expected that different tastes will generate local branding effects which mitigate against substitutability and create differences in hedonic prices. Nevertheless, two locations with both similar characteristics and similar hedonic prices must be good substitutes, as it will be very difficult to distinguish between them. Therefore, simultaneous similarity in hedonic characteristics and their shadow prices may be a sufficient condition for substitutability. However, this is not a necessary condition, because two types of houses with very different hedonic characteristics can be good substitutes. For example, a flat in a central location can be an alternative to a more peripheral detached house with a similar price; hence, proximity in location is also required.

The conceptual notion behind spatial submarkets discussed above implies that the price determining (hedonic) mechanism can be heterogeneous over space. This spatial heterogeneity, reflecting the absence of a single equilibrium in the housing market, can originate from demand and supply side factors, institutional barriers or discrimination, each of which can cause differentials across neighbourhoods in the way housing attributes are valued by consumers and house prices determined (Anselin *et al.*, 2010). However, if spatial submarkets exist and are ignored, an average price across all the territory is estimated that ignores submarket heterogeneity.

The classical urban model in the Alonso-Muth-Mills tradition predicts a decrease in prices with distance from the city centre, though there may be spatial variation in relative preference for centrality itself. Other models based on localised amenities or multiple centres imply a stronger impact of access to local amenities. Like distances,

[1] The underlying assumption is that the time trend is the sole latent factor, inclusion of which ensures that the spatial structure contains only spatial weak dependence (Pesaran and Tosetti, 2011). Pryce (2013) does not explicitly state this assumption, but it is implied by the methodology, based on computation of inflation in house prices at different locations. This assumes such an underlying spatial model, together with the assumption that inclusion of the time trend ensures spatial weak dependence.

the implicit prices for dwelling characteristics and size may also vary spatially, reflecting either supply constraints or residential sorting. Adair *et al.* (1996) and Malpezzi (2003), among others, discuss intra-urban variation in the price of housing amenities using hedonic models. Thus, heterogeneity is a key element of housing markets and its disregard seriously affects the understanding of market diversification. Also, it is likely to bias average price estimates because the error term of the regression model may be correlated with the included regressors.

There are two main methods to model spatial heterogeneity. First, one may allow coefficients in the hedonic model to vary across submarkets, and use the estimated variation to infer on residential neighbourhood choice and urban spatial structure. The second method, geographically weighted regressions (GWR) (Fotheringham *et al.*, 1998), is a form of locally weighted regression that we discuss later in the paper (section 4.2.1).

2.2.3. Spatial dependence and spatial weights matrix

Spatial dependence arising from spatial spillovers or contagion effects leads to spatial autocorrelation, implying that prices of nearby houses or related submarkets tend to be more similar. Spatial autocorrelation can also result from incorrectly modelled spatial heterogeneity, measurement errors in regressors, omitted variables or unmodelled spatial patterns in hedonic features (Anselin, 1988; LeSage and Pace, 2009).

Spatial dependence is very common in housing markets, and a feature that we use in this paper to develop inferences for a functional regression model. The recent literature has discussed bias and loss of efficiency that can result when spatial effects are ignored in the estimation of hedonic models. The use of spatial econometric models to address spatial autocorrelation is becoming increasingly standard (Anselin and Lozano-Gracia, 2008; LeSage and Pace, 2009; Anselin *et al.*, 2010). The usual approach to the representation of spatial interactions is to define a spatial weights matrix, denoted W, which represents a theoretical and *a priori* characterisation of the nature and strength of spatial interactions between different submarkets or dwellings.[2] These spatial weights represent patterns of diffusion of prices and unobservables over space, and thereby provide a meaningful and easily interpretable representation of spatial interaction (spatial autocorrelation). Given a particular choice of the spatial weights matrix, there are two important and distinct ways in which spatial dependence is modelled – the spatial lag model and the spatial error model. In the former, the hedonic regression includes as an additional regressor – the spatial lag of the dependent variable y (which in this case is price), represented by Wy:

[2] For a setting with n spatial units, W is an $n \times n$ matrix with zero diagonal elements. The off-diagonal elements are typically either dummy variables for contiguity or inversely proportional to distance between a pair of units.

$$\underline{y} = \rho W \underline{y} + X\beta + \underline{\varepsilon}, \quad (2)$$

where *X* denotes hedonic characteristics and the regression errors (ε) are idiosyncratic. By contrast, in the spatial error model, the regression errors are spatially dependent on their spatial lag, $W\varepsilon$:

$$\underline{y} = X\beta + \underline{\varepsilon}, \ \underline{\varepsilon} = \lambda W \underline{\varepsilon} + \underline{\eta}. \quad (3)$$

There is a large literature on estimation and inferences for these two models. In the spatial lag model, the endogenous spatial lag implies that OLS estimates are biased, while in the spatial error model, they will be unbiased but inefficient. The spatial weights are typically modelled either by spatial contiguity, or as functions of geographic or economic distance. The distance between two spatial units reflects their proximity with respect to prices or unobservables, and hence the spatial interaction between a set of units can be measured by a function of the distance between them. However, spatial data may be anisotropic, where spatial autocorrelation is a function of both distance and the direction separating points in space (Gillen *et al.*, 2001). Similarly, spatial interactions may be driven by other factors, such as trade weights, transport cost, travel time, and socio-cultural distances. The choice typically differs widely across applications, depending not only on the specific economic context but also on availability of data; for extensive discussion, see Bhattacharjee and Jensen-Butler (2013). Most studies place emphasis on either spatial heterogeneity or spatial dependence but not both. Bhattacharjee *et al.* (2012) developed a framework that emphasizes all the three distinct but interconnected features of space – spatial heterogeneity, spatial dependence and spatial scale.

Finally, while the traditional literature is based on an *a priori* known structure of spatial dependence, or a spatial weights matrix *W*, and then examined spatial dependence and spatial heterogeneity implied by this *W*, a branch of the current literature treats these weights as unknown. Based on a given definition of urban submarkets (or a fixed set of spatial locations) and panel data on these spatial units, Bhattacharjee and Holly (2013) and Bhattacharjee and Jensen-Butler (2013) proposed several methods to estimate the spatial weights matrix between the submarkets. Bhattacharjee *et al.* (2012) extended the methodology to a purely cross-section setting, where the delineation of submarkets was assumed known *a priori*. By contrast, this paper focuses on identifying submarkets in a setting where *W* may be known, or even unknown and potentially endogenous.

3. Delineation of Housing Submarkets

Dividing a large market into submarkets raises numerous theoretical and methodological questions (Rothenberg

et al., 1991). Theoretically, a submarket corresponds to a local equilibrium between supply and demand. However, the way submarkets are delineated depends on the level of aggregation and methods for clustering basic spatial units into submarkets. In practice, ad-hoc methods based on predefined geographical boundaries are often used; sometimes regions defined *a priori* are statistically tested for distinctness (Bourassa *et al.*, 1999).

3.1. Submarkets based on similarity in hedonic characteristics and prices

A branch of the urban studies literature has focused on more systematic methods for defining submarkets. A common approach is to proceed first by conducting principal component or statistical factor analysis on a large number of hedonic characteristics of houses to extract a small number of meaningful factors. Next, clustering methods are used to obtain a set of submarkets that maximise the degree of internal (within-submarket) homogeneity and external (across submarket) heterogeneity; see Bourassa *et al.* (1999) for further discussion. This is related to the definition of submarkets by similarity of hedonic housing features (Rothenberg *et al.*, 1991).

Another approach is based on homogeneous hedonic prices, using as a measure of homogeneity small residuals from a hedonic pricing model estimated separately for each submarket (Bourassa *et al.*, 1999, 2003). The objective is to use submarkets to improve accuracy of hedonic predictions for mass appraisal purposes. Further, homogeneity in hedonic prices is deeply rooted in the basic concepts which underlie hedonic models.

3.2. Submarkets based on substitutability

The above approaches are not entirely satisfactory from a housing economics point of view. They do not pay explicit attention to the demand side of the housing market, which is where individual households make neighbourhood and housing choice decisions. Similarity in hedonic housing characteristics relate to the supply side, and similarity in hedonic prices relate to market equilibria which is the outcome of demand and supply sides of submarkets. The concept of substitutability is useful to the extent that it can be interpreted as reflecting the synthetic valuation of houses by buyers, and therefore offers understanding of the demand side (Pryce, 2013).

Grigsby *et al.* (1987) defined submarket as a region where the dwellings are reasonably close substitutes, but relatively poor substitutes for dwellings in other submarkets, and Pryce (2013) proposed submarket delineation by taking house prices as the determinant of housing choice[3] and by evaluating the cross-price elasticity of price for each pair of housing properties. Two houses are deemed to lie within the same submarket if this cross-price

[3] Taking house prices, rather than hedonic characteristics, as the sole basis for evaluation of substitutability is not an innocuous modelling assumption. See Pryce (2013) for further discussion.

elasticity is close to unity, implying therefore that they are substitutable. Thus, Pryce (2013) uses house price inflation for computation of the elasticities. Placing the above methodology within the context of a structural spatial econometric model is useful for our discussion. In essence, Pryce (2013) assumes a spatial error model:

$$\underline{y}_t = \underline{y}_{t-1} + \underline{\varepsilon}_t,\ \underline{\varepsilon}_t = W\underline{\varepsilon}_t + \underline{\eta}_t, \tag{4}$$

where $\underline{y}_t$ denotes the vector of prices (in logarithms) across all houses, and $\underline{y}_{t-1}$ its lagged value, so that $\underline{\varepsilon}_t$ denotes the growth rate. Then, the elements of W are the cross-price elasticities for each pair of houses.[4] The model does not include any regressors other than lagged logarithm of prices with unit coefficient, which is assumed to ensure that the regression errors (growth rates) are stationary in the temporal domain. Although simplistic, the model itself is structural because it assumes that the process of diffusion of shocks (η_t) is driven by an underlying spatial structure in W. Elements of W are estimated and then used for delineation of submarkets.

From a spatial econometric point of view, this approach deals with potential temporal nonstationarity by inclusion of the lag on the right hand side. This also suggests a natural interpretation of elasticity as a cause-effect relationship over time. However, strong spatial dependence is a potential problem. This would be evident if some elements of W are close to unity (or even larger), which would imply violation of the spatial granularity condition (Pesaran and Tosetti, 2011). Further, violation of this condition is expected because cross-price elasticities are by definition close to unity for houses within the same submarket. One can ensure that cross submarket spatial diffusion is bounded, and therefore the spillover of house price shocks across the submarkets is spatially stationary. However, spatial weights of houses within the same submarket will be large. Therefore, without suitable modifications, model (4) cannot be cast into the framework of contemporary spatial econometrics.

3.3. Submarkets based on a structural spatial lag model

The above discussion suggests indicates that the model (4) may be extended in two ways. First, violation of the spatial granularity condition points towards spatial strong dependence, which is caused by ignoring the effect of common factors (Pesaran, 2006; Pesaran and Tosetti, 2011). The solution is to include regressors that will take strong dependence out of the model; see Bhattacharjee and Holly (2013) for further discussion. In the current

[4] Since $\underline{\varepsilon}_t$ denotes the growth rate of prices, the spatial error part of (4) models how growth rate in a location is related linearly to the growth rates at all other locations. The elements of W are the corresponding coefficients, or cross-price elasticities. In Pryce (2013), two houses are viewed as being substitutable if the cross-price elasticity is close to unity, which in Equation (4) implies the corresponding spatial weights are close to unity.

context, hedonic characteristics can be added to the model, allowing the cross-price elasticities to be measured more robustly. Second, assumption of a spatial error model is somewhat simplistic. The common belief is that house prices are spatially endogenously determined by the interaction between housing choices of economic agents; hence the spatial lag model (2) is more appropriate and allows stronger structural interpretations.

In a setting where *W* is unknown (Bhattacharjee and Holly, 2013; Bhattacharjee and Jensen-Butler, 2013), *W* and *ρ* are not separately identifiable. Hence, we assume without loss of generality that *ρ = 1*. Further, an unknown *W* is not in general identified, and structural assumptions are required for identification. Following Bhattacharjee and Jensen-Butler (2013), we make the assumption that *W* is symmetric. In applications, spatial weights matrix *W* is often based on distances, which are symmetric by definition.

Based on the above discussion, we make the following assumption.

Assumption 1. ***Spatial lag model.*** *The dependent variable y follows a spatial lag model*

$$y = Wy + X\beta + \varepsilon \Rightarrow y = (I - W)^{-1} X\beta + (I - W)^{-1}\varepsilon. \tag{5}$$

with full spatial heterogeneity in both the slope and intercept (heterogeneity in β across the territory, plus location fixed effects). W is unknown but symmetric, and satisfies the spatial granularity condition $\rho(W) < 1$*, where* $\rho(W) = \max\left\{\|W\|_1, \|W\|_\infty\right\}$ *is the norm of W,* $\|W\|_1 = \max_{1\le j\le n} \sum_{i=1}^{n} |w_{ij}|$ *the column norm of W, and* $\|W\|_\infty = \max_{1\le i\le n} \sum_{j=1}^{n} |w_{ij}|$ *the row norm of W. The regression errors,* ε*, have mean zero, and X are regressors that are uncorrelated with* ε *and are not collinear, that is, have a positive definite covariance matrix.*

The spatial granularity condition implies that there is no spatial strong dependence (Pesaran and Tosetti, 2011). If there are latent factors that can cause violation of the condition, they are included as regressors in model (5).

As an illustration, consider a simple spatial lag model regressing logarithm of price per square meter (*y*) on logarithm of living space (*x*), allowing for spatial heterogeneity and endogenous spatial dependence. Further, to fix ideas, let us first consider a sample of only two locations with potentially different slopes, with only one house in each location. Then:

$$\begin{aligned} &\begin{pmatrix} y_1 \\ y_2 \end{pmatrix} = W\begin{pmatrix} y_1 \\ y_2 \end{pmatrix} + \begin{pmatrix} x_1\beta_1 \\ x_2\beta_2 \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \end{pmatrix}, \quad W = \begin{bmatrix} 0 & w_{12} \\ w_{21} & 0 \end{bmatrix} \\ &\Rightarrow \begin{pmatrix} y_1 \\ y_2 \end{pmatrix} = [I - W]^{-1}\begin{pmatrix} x_1\beta_1 \\ x_2\beta_2 \end{pmatrix} + [I - W]^{-1}\begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \end{pmatrix} \approx [I + W]\begin{pmatrix} x_1\beta_1 \\ x_2\beta_2 \end{pmatrix} + [I + W]\begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \end{pmatrix}, \end{aligned} \tag{6}$$

where the final step follows because the spatial weights are very small compared to unity, so that $[I-W]^{-1} \approx [I+W]$. This assumption is valid under weak spatial dependence, in which case the spatial granularity condition in Assumption 1 holds.

Equation (6) emphasizes that both y_1 and y_2 are endogenously determined by each other, and in addition are functions of x_1, x_2, β_1 and β_2. Further, the response at location i, y_i, is a function of the regressor at the same location (x_i) with slope β_i, but also the regressor at every other location j, x_j, but with a different slope ($w_{ij}\,\beta_j$). Conceptually, y_1, y_2, x_1, x_2, β_1 and β_2 are thought of as functions of space, where $s \in S$ is a representative point in the spatial domain. We assume that these are smooth functions so that all partial derivatives are well defined.

Assumption 2. ***Smoothness.*** *The functional regression coefficient, β(s) varies smoothly over the compact set S. That is, β(s) has derivatives at every s ϵ S. Likewise, the functional random variables x(s) and y(s) have mean functions,* $\overline{X}(s)$ *and* $\overline{Y}(s)$ *respectively, that are smoothly varying over S.*

By Assumption 2, the partial derivatives $\partial\beta/\partial s$, $\partial x/\partial s$ and $\partial y/\partial s$ are well defined. Then, we have the following result, where all partial derivatives are interpreted with respect to space, $s \in S$.

Theorem 1: *Under Assumptions 1 and 2, two houses are substitutable, that is the cross-price elasticity of price is close to unity, if their hedonic characteristics, prices and location are similar.*

Proof: *By the granularity condition in Assumption 1, the elements of W are small, and hence up to first order Taylor expansion,* $(I-W)^{-1} \approx (I+W)$*. The idiosyncratic errors can be ignored in computation of cross-price elasticities. Then, for any two distinct houses in locations i and j:*

$$\frac{\partial y_i}{\partial y_j} \approx \frac{1+w_{ji}}{1+w_{ij}} \cdot \frac{\partial(x_i\beta_i)}{\partial(x_j\beta_j)} = \frac{1+w_{ji}}{1+w_{ij}} \cdot \frac{x_i d\beta_i + \beta_i dx_i}{x_j d\beta_j + \beta_j dx_j} = \frac{x_i d\beta_i + \beta_i dx_i}{x_j d\beta_j + \beta_j dx_j}, \tag{7}$$

where $(1+w_{ji})/(1+w_{ij}) = 1$ *since the spatial weights are symmetric (Assumption 1). Thus, sufficient conditions for houses i and j to be (approximately) substitutable are: (i)* $x_i \approx x_j$*; (ii)* $\beta_i \approx \beta_j$*; and (iii) the locations i and j are in each other's neighbourhood, so that by Assumption 2,* $d\beta_i \approx d\beta_j$ *and* $dx_i \approx dx_j$*. The proof in the general case follows by noting that elasticities involve only a pairwise comparison between 2 properties i and j, and other houses can be ignored because elements of W are small.*

Theorem 1 has important implications for delineation of submarkets. First, a sufficient condition for (houses in) locations *i* and *j* to be substitutable is that the spatially varying *x* and *β*'s in the two locations match, and their

slopes match as well. This implies spatial clustering of the *x*'s and the *β*'s. In other words, based on the close substitutability definition, submarkets may be delineated by spatial clustering jointly on both these two dimensions. Second, the insights can be easily extended to the case of multiple regressors (or hedonic factors). Here, clustering should include all the included hedonic factors as well as their spatially varying slopes. Third, the method in Pryce (2013) is appropriate if there are no regressors. In this case, the cross-price elasticities will be solely determined by elements of *W*. However, because of spatial nonstationarity, the model will then not offer any useful structural interpretation, and the estimates of elasticities are also likely to be biased.

4. A Synthesis of Empirical Approaches

Next, we consider implementation of a procedure for delineating submarkets. For this purpose, we develop a synthesis of several empirical approaches rather than a purely spatial econometric framework.

4.1. The limits of spatial econometrics?

Recent spatial econometrics literature has considered estimation of spatial weights (Bhattacharjee and Holly, 2013; Bhattacharjee and Jensen-Butler, 2013; Bailey *et al.*, 2014), endogenous spatial structure (Kelejian and Piras, 2014), and connections between different aspects of space (Bhattacharjee *et al.*, 2012). There are, however, two leading aspects where the framework needs to be extended.

First, while the above framework uniquely combines spatial heterogeneity and spatial dependence, the way spatial dependence is modelled is somewhat unsatisfactory. Specifically, in restricting spatial spillovers to a spatial error model, adequate attention is not paid to endogenous evolution of space itself. At the same time, it is perhaps inevitable that housing markets are endogenously related over space. Location choices and consequently prices are not only spatially contingent, but also endogenously connected, which implies that spatial dependence through a spatial lag model is more appropriate. While the literature has paid elaborate attention to spatial lag dependence, for example, Anselin and Lozano-Gracia (2008) and Anselin *et al.* (2010), this has been in a context where the spatial weights are known *a priori*, and there is no spatial heterogeneity.

Second, the above framework assumes a segmentation into housing submarkets which is given *a priori*. We need delineation of submarkets based on hedonic characteristics and prices that are spatially heterogeneous within a spatial context where spatial dependence is endogenous.

4.2. Some alternate approaches

In summary, a new framework is required. We now turn to alternative perspectives from the geography and

statistics literatures, specifically local regressions (for example, geographically weighted regressions, GWR) and functional data analysis (FDA).

4.2.1. Geographically (or locally) weighted regression

In the literature, spatial heterogeneity is typically modelled using locally weighted regressions (McMillen, 1996), of which the Geographically Weighted Regression (GWR) (Fotheringham *et al.*, 1998) is perhaps the most popular. GWR is a nonparametric regression method that replaces the single regression coefficient in a linear model with a series of (geographically weighted) estimates:

$$E[\int_S Y(s) f_{h,i}(s) ds] = \alpha_i + \beta_i \int_S X(s) f_{h,i}(s) ds, \quad (8)$$

where Y and X are both defined over a territory S determined by a medium or large urban housing market, i is a location within the spatial domain S, $f_{h,i}(s)$ is a kernel density with bandwidth h and centred on location i, the regression slope β_i varies over space, and α_i can be interpreted as a location specific fixed effect. In effect, this method provides pointwise estimates β_i of the regression effect of a kernel weighted local average of Y on a similarly kernel weighted local average of X.

4.2.2. Functional data analysis

Functional data analysis (FDA) is a framework and collection of tools for statistical analysis of functional data, which refers to curves, surfaces or anything else that varies over a continuum; see, for example Ramsay and Silverman (2005). The main challenge in FDA is that functional data (curves or surfaces) are intrinsically infinite dimensional while sample sizes are limited. Hence the data have to be projected on the span of a suitable basis, assuming that the data are intrinsically smooth, while observed data include measurement error. In the typical case where the functional domain is time, inferences can be based on a Fourier basis for periodic data or smoothing splines for data that are not periodic (Ramsay and Silverman, 2005).

In our spatial context, the functional linear regression model takes the form:

$$E[y_i] = \alpha + \int_S \beta(s) x_i(s), \quad (9)$$

where the response (y) is scalar, and the regressor (x) and slope (β) are functional. Application of FDA in the spatial domain is not straight forward. Unlike time series, there is no well-defined ordering of spatial observations, and neither a direction of information flow. Hence the choice of a basis space is challenging. Guillas and Lai

(2010) proposed bivariate splines over triangulations. However, this does not take into account the spatial context in terms of the geography of the region or spatial dependence. Here, we adapt the intuitive and powerful functional principal components estimator (Cai and Hall, 2006; Hall and Horowitz, 2007) to our spatial context.

4.3. A Proposed Synthesis of Different Perspectives

Thus, we propose a new framework, based on a synthesis of spatial econometrics and functional data analysis. This framework addresses some of the limitations of the previous approaches. Intuition suggests that such a synthesis may be promising. For illustration, consider again the simple spatial lag model in (6) specific to two housing properties, but incorporating heterogeneity in slopes. As discussed above, the reduced form

$$\begin{pmatrix} y_i \\ y_j \end{pmatrix} \approx [I + W] \begin{pmatrix} x_i \beta_i \\ x_j \beta_j \end{pmatrix} + [I + W] \begin{pmatrix} \varepsilon_i \\ \varepsilon_j \end{pmatrix} \tag{10}$$

implies a regression model where, in addition to $x_i\beta_i$, the right hand side also includes $x_j\beta_j$, but with a much smaller weight, since $w_{ji} << 1$. This suggests a functional regression model where the response variable is scalar and the functional regressor is $x_i(s) = x_i\ f_{h,i}(s)$ with kernel weights $f_{h,i}(s)$ proportional to the elements of $[I - W]^{-1} \approx [I + W]$. This intuition generalises to multiple locations and houses.

Thus, the spatial lag model is a special case of the functional regression model, corresponding to a specific definition of the functional regressor. Further, as the bandwidth h reduces to zero, GWR and functional regression becomes very similar. This suggests that a synthesis of perspectives from spatial econometrics, GWR and FDA may deliver a spatial statistical model that is very rich and enable the full range of spatial analyses of housing markets. Importantly, the model offers efficiency and robustness by using information from neighbours through the spatial weights matrix.

The above model addresses both the limitations of the previous approaches. First, the proposed framework combines the regressor (x_i) and (kernel) spatial weights, $f_{h,i}(s)$, into a functional regressor, $x_i(s)$. This allows inference on a functional slope to proceed beyond the limitations of exogenously specified submarkets or spatial weights. Now, endogeneity in the spatial weights can be modelled in conventional ways. One would need either a dynamic model for how these weights evolve over time, or use suitable instruments for $x_i(s)$.

Second, the framework allows submarkets to evolve endogenously without the need to delineate housing submarkets *a priori*. As discussed before, the literature has defined submarkets either as a collection of locations

that have close hedonic substitutability (Rothenberg *et al.*, 1991; Bourassa *et al.*, 2003), or by the degree of substitutability (Grigsby *et al.*, 1987; Pryce, 2013). In the context of a hedonic model with homogenous slopes, the two definitions are equivalent. However, this is not true when there is heterogeneity across submarkets. This heterogeneity relates not only to the hedonic characters, but also the shadow prices of such features. As our Theorem 1 suggests, in the presence of such heterogeneity, housing submarkets should be delineated by spatial clustering jointly of the functional partial effect $\beta(s)$ and the functional surface of the hedonic characteristics $x(s)$.

5. Methodology

Our methodology for delineation of submarkets starts with estimating a functional regression model

$$E[y_i] = \alpha(s) + \int_S \beta(s) x_i(s), \quad x_i(s) = x(s) f_{h,i}(s), \tag{11}$$

where y_i is a scalar response at location i, $x(s)$ is defined over a spatial domain S corresponding to an urban housing market (and, unlike the typical functional regression model, not a subset of the positive real line $\mathbf{R}^+$), and $f_{h,i}(s)$ is a kernel density with bandwidth h and centred on location i.

5.1. Estimating the Functional Regression Model

The functional linear regression model (11) is based on a large (and potentially infinite) dimensional functional regressor that needs to be regularised. Hence, estimation involves projection to a suitable basis space. We consider the functional principal components estimator (Hall and Horowitz, 2007). However, application of the method presents some challenges in our setting. For a specific location i, the functional surface of $x_i(s)$ is a weighted form of x_i, with the weights given by a kernel $f_{h,i}(s)$. This kernel places a large weight in the neighbourhood of location i, but relatively small weights elsewhere. This implies that the functional surface has very sparse information which in turn requires a large number of principal components and also produces a poor approximation. For this problem of regularisation, we develop a variant of functional principal components.

Our data generating process is as follows. The data constitute a collection of dependent pairs $(X_1,Y_1), (X_2,Y_2),\ldots,(X_n,Y_n)$ indexed on n locations in a compact set $S \subset \mathbf{R}^2$. For a specific location $i \in S$, both Y and X are scalar random variables. The Y_i are generated by a functional linear regression model

$$Y_i = \alpha + \int_S \beta X_i^* + \varepsilon_i, \qquad i = 1,\ldots n,$$

$$X_i^*(u) = \begin{cases} X_i & \text{if} \quad u = i \\ X_j f_{ij} & \text{if} \quad u = j \in \{1,\ldots,n\}, i \neq j, f_{ij} = f_{h,i}(j) \\ 0 & \text{otherwise.} \end{cases} \tag{12}$$

The errors ε_i are potentially spatially dependent, but are identically distributed with finite variance and zero mean, and are independent of the explanatory variables; no distributional assumptions are made.

The main issue with model (12) is that the functional regressor surface of X_i^* is very irregular. By assumption 2, the mean of the underlying regressor, $\overline{X}(u)$, varies smoothly over *S*. However, the combination of a large (unit) weight at location *i* with a kernel function elsewhere renders X_i^* very irregular and spiky. Hence, usual regularisation by principal components as in Cai and Hall (2006) and Hall and Horowitz (2007) is not feasible. The tuning parameter (number of principal components) will be very large, and correspondingly, the spacings between eigenvalues are very small, so that the results in Hall and Horowitz (2007) are not directly applicable.

Hence, our approach focuses on directly regularising the surface of $\overline{X}(u)$ using functional principal components. To motivate the approach, consider the surface of the functional regressor X_i^* for the specific observation *i*. The challenge here is the spiky nature of X_i^*, due to a very large (unit) weight at the location of observation *i*, together with much lower weights (f_{ij}) at other locations. Our object of inference here is the functional surface of the regression coefficient (β) which is smooth (Assumption 2). Hence, the regressor at this location can be potentially combined with values in its neighbourhood. By averaging, the irregular functional regressor surface can be smoothed. This suggests that partitioning *S* into several (*K*) regions (say, $\{P_1, P_2, \ldots, P_K\}$) may be a good starting point. This may also be viewed as a first stage of regularisation, where the basis function is a histogram sieve.

Then, we apply functional principal components to the averaged regressor process across the partitions, that is to ($\bar{x}_1, \bar{x}_2, \ldots, \bar{x}_K$), where $\bar{x}_k = E(X_i \mid i \in P_k), k = 1, \ldots, K.$ The procedure poses two major challenges: (a) by averaging, we would lose variability across observations, and therefore implementation of functional principal components is challenging; and (b) if we were to implement principal components, we need to develop a method similar to Hall and Horowitz (2007) to then use these principal components to estimate the functional surface of the regression coefficient (β).

For (a), the same spike that was a problem earlier now helps once a histogram sieve (partition) has been placed. Consider the compact space *S* partitioned into *K* regions $P_1, P_2, \ldots, P_K$, with corresponding sample sizes $n_1, n_2, \ldots, n_K$, with $\sum n_k = n.$ For notational simplicity, we denote by *k(i)=k* the partition that observation *i*

belongs to, that is $i \in P_k$. Then, the sieve functional regressor for observation i is

$$[n_1 f_{i1}\bar{x}_1, n_2 f_{i2}\bar{x}_2, \ldots, n_{k-1} f_{i,k-1}\bar{x}_{k-1}, [(1-f_{ii})X_i + n_i f_{ii}\bar{x}_k], n_{k+1} f_{i,k+1}\bar{x}_{k+1}, \ldots, n_K f_{iK}\bar{x}_K]. \quad (13)$$

We divide the j-th element of the functional regressor vector (13) by the scalar exogenous weight $n_j f_{ij}$,:

$$X_i^{**} = \left[\bar{x}_1, \bar{x}_2, \ldots, \bar{x}_{i-1}, \bar{x}_i + \frac{1-f_{ii}}{n_i f_{ii}} X_i, \bar{x}_{i+1}, \ldots, \bar{x}_K\right]. \quad (14)$$

Now, there is variation in the functional regressor surface across observations within each partition, and hence functional principal components can be implemented. At the same time, $\frac{X_i - f_{ii}\bar{x}_i}{n_i f_{ii}} \to 0$ as $n \to \infty$, so that in large samples, (14) approximates the average process. Thus it is expected to be smooth over space since, by Assumption 2, the functional surface of the average $\bar{X}(s)$ is smooth. In large samples, when variation in X_i does not matter for the construction of the functional regressor, $X^* \approx \bar{X}Z$, where $\bar{X} = [\bar{x}_1, \bar{x}_2, \ldots, \bar{x}_K]$ and Z is a vector with value 1 at location i and 0 otherwise.

With (b) note that, for the data within a specific partition P_k, the functional regression coefficient for the partition is β_k times $n_k f_{kk}$, where the coefficient itself corresponds to the k-th element of X^{**}. Note also that, within this same partition, there is no cross-section variation in the other elements of X^{**}, and hence their effects are encompassed within a fixed effect for the partition. Hence, the functional surface of the regression coefficient can be estimated by a functional regression model where the dependent variable is measured in deviations from the local (within partition) mean, and the functional regressor is given by equation (14).

In our application, we have a finite but large-dimensional setting where the number of partitions (K) is large. Below, we assume that the spatial design, given by Z, is held fixed in repeated sampling. Finally, we obtain our estimator $\hat{\beta}$ by dividing the k-th element of the functional regression estimator by the deterministic scalar $n_k f_{kk}$.

Thus, consider the modified linear functional regression model:

$$Y_i = a + \int_S b(u) X_i^{**}(u) du + \varepsilon_i, \qquad i = 1, \ldots n,$$

$$X_i^{**}(u) = \begin{cases} \bar{X}(u) + \dfrac{1-f_{ii}}{n_i f_{ii}} X_i & \text{if} \qquad u = i \\ \bar{X}(u) & \text{otherwise.} \end{cases} \quad (15)$$

The X_i^{**} are random functions on $S \subset \mathbf{R}^2$, the intercept *a* and the errors ε_i are scalars and the slope *b*, our main object of inference, is a function on *S*. Let (X^{**}, *Y*, ε) denote a generic (X_i^{**} , Y_i, ε_i). Define

$$K(u,v) = \operatorname{cov}\{X^{**}(u), X^{**}(v)\}, \quad \hat{K}(u,v) = \frac{1}{n}\sum_{i=1}^{n}\{X_i^{**}(u) - \overline{X}^{**}(u)\}\{X_i^{**}(v) - \overline{X}^{**}(v)\}, \quad u,v \in S,$$

where $\overline{X}^{**}(.) = n^{-1}\sum_i X_i^{**}(.)$. Write the spectral expansions of *K* and $\hat{K}$ as:

$$K(u,v) = \sum_{j=1}^{\infty}\kappa_j\phi_j(u)\phi_j(v), \qquad \hat{K}(u,v) = \sum_{j=1}^{\infty}\hat{\kappa}_j\hat{\phi}_j(u)\hat{\phi}_j(v),$$

where $\kappa_1 > \kappa_2 > \ldots > 0$ and $\phi_1, \phi_2, \ldots$ are the eigenvalue and corresponding orthonormal eigenvector sequences of the linear operator with kernel *K*, and similarly $\hat{\kappa}_1 \geq \hat{\kappa}_2 \geq \ldots \geq 0$ and $\hat{\phi}_1, \hat{\phi}_2, \ldots$ for the kernel $\hat{K}$. The sequences $(\hat{\kappa}_j, \hat{\phi}_j)$ of eigenvalues and eigenvectors of the empirical covariance matrix of X^{**} constitute an estimator of (κ_j, ϕ_j). Then, the functional principal components estimator (Hall and Horowitz, 2007) of the regression slope the slope *b(.)* is given by

$$\hat{b}(u) = \sum_{j=1}^{m}\hat{b}_j\hat{\phi}_j(u), \tag{16}$$

where the spectral cutoff *m* is a tuning parameter, $\hat{b}_j = \hat{\kappa}_j^{-1}\hat{g}_j$, $\hat{g}_j = \int \hat{g}\hat{\phi}_j$, and $\hat{g}(u) = \frac{1}{n}\sum_{i=1}^{n}\{Y_i - \overline{Y}\}\{X_i^{**}(u) - \overline{X}^{**}(u)\}$.

Note that the functional regression estimator (16) is a least squares estimator, depending only on the sample covariance function of the functional regressor X^{**}, truncated at a finite cutoff for the spectral expansion, and the covariance function of *Y* and X^{**}. Thus, it is essentially a method of moments estimator that requires neither independent errors nor a specific error distribution, but is based on mean zero errors and orthogonality of the regressor and the errors. This is useful in our context, since the errors in our reduced form spatial model (7) are correlated. Further, the functional regressor in our spatial setting can be endogenous, either because we use estimated spatial weights, or because the underlying regressor *X* or the weights matrix *W* are endogenous. In such cases, an instrumental variables estimator can be constructed. Finally, note that, in our setting, *Y* is measured in terms of deviation from the local (within partition) mean, thus allowing for spatial fixed effects.

Next, we make assumptions required for consistency and convergence rates of our estimator.

Assumption 3: ***Technical assumptions for functional regression inference.***

(a) *The data are generated by fixed spatial design, so that Z is not stochastic.*

(b) *All other technical assumptions in Hall and Horowitz (2007) hold. Specifically, conditions on the distribution of X^{**}, distribution of ε, eigenvalues and Fourier coefficients hold.*

i) X has finite fourth moments, and hence so does X^{**}. The error ε_i are identically distributed with zero mean and finite variance not exceeding some constant C.

ii) *Consider the Karhunen-Loève expansion of the random function X^{**}: $X^{**} - E(X^{**}) = \sum_{j=1}^{\infty} \xi_j \phi_j$, where the ξ_j are pairwise uncorrelated zero mean random variables with variances κ_j that are eigenvalues of the expansion. The κ_j satisfy the spacing condition $\kappa_j - \kappa_{j+1} \geq C^{-1} j^{-\alpha-1}$ for all j and some exponent α > 1.*

iii) *Let $b_j = \kappa_j^{-1} g_j$ where $g(u) = E\left[(Y - EY)(X(u) - EX(u))\right]$ and $g_j = \int g \phi_j$. The b_j's satisfy $|b_j| \leq C j^{-\delta}, \delta > \frac{1}{2}\alpha + 1$.*

iv) *The tuning parameter m increases with n such that $m / n^{1/(\alpha+2\delta)}$ is bounded away from zero and infinity.*

Then, the functional surface *b(s)* can be estimated by the functional principal components estimator in Hall and Horowitz (2007). However, our object of inference is the functional surface of *β(s)* in (12) and not the *b(s)* in (15). Assumption 3(a) provides a simple way to go from the $\hat{b}(s)$, estimated by functional principal components as in (16), to the $\hat{\beta}(s)$. In the finite but large dimensional setting, or when a histogram sieve is placed on the spatial domain, we simply have $\hat{\beta}(s) = \hat{b}(s)/(n_k f_{kk}), s \in P_k$. Assumptions 3(b) are discussed in Hall and Horowitz (2007). Condition 3(b)i) is standard. Condition (b)ii) ensures that all eigenvalues have unit multiplicity, and their spacing decreases exponentially, so that we need a small smoothing spectral cutoff. Assumption (b)iii) ensures that the Fourier coefficients are bounded below and above. Condition (b)iv) ensures that the number of basis function terms used in the smoothing process of *b* is much smaller than *n*. Then, we have the following result.

Theorem 2 (Hall and Horowitz, 2007): *Let $\Im(C, \alpha, \delta)$ denote the set of distributions F of (X^{**}, Y) that satisfy Assumption 3 for given values of C, α and δ. Let B denote a class of measurable functions $\bar{b}$ of the data*

$\left(X_1^{**},Y_1\right),\ldots,\left(X_n^{**},Y_n\right)$ *generated by (15). Then,* $\hat{b}(s)\xrightarrow{P}b(s)$. *Specifically,*

$$\lim_{D\to\infty}\limsup_{n\to\infty}\sup_{F\in\Im}P_F\left\{\int_S\left(\hat{b}-b\right)^2>Dn^{-(2\delta-1)/(\alpha+2\delta)}\right\}=0\quad\text{as}\quad n\to\infty$$

and

$$\liminf_{n\to\infty}n^{(2\delta-1)/(\alpha+2\delta)}\inf_{\bar{b}\in\mathrm{B}}\sup_{F\in\Im}\int_S E_F\left(\bar{b}-b\right)^2>0,$$

which then imply that for each $F\in\Im$, $\int_S\left(\hat{b}-b\right)^2=O_p\left(n^{-(2\delta-1)/(\alpha+2\delta)}\right)$.

For the technical details of the proof and associated discussion, refer to Hall and Horowitz (2007). The functional principal components estimator is a method of moments estimator. The identical distribution condition for the errors is stated in Assumption 3(b)i), but is not required in the proof of Theorem 2 beyond the zero covariance between X^{**} and ε. The technique of proof is somewhat nonstandard, in showing that the supremum and infemum have the same rate of convergence, and moreover in using probability (P_F) rather than expectation (E_F) in the supremum statement. The rate of convergence $n^{-(2\delta-1)/(\alpha+2\delta)}$ is generic to noisy inverse problems. The main result was shown in Hall and Horowitz (2007). Our main innovation here is to adapt the above general result to an irregular (spiky) functional regressor surface. We achieve this by using a histogram sieve.

Corollary 1: *Under Assumption 3,* $\hat{\beta}(s)\xrightarrow{P}\beta(s)$ *and for each* $F\in\Im$,

$$\int_S\left(\hat{\beta}-\beta\right)^2=O_p\left(n^{-(2\delta-1)/(\alpha+2\delta)}\right).$$

Proof: *The proof follows directly from Theorem 1, noting that by Assumption 3(a),* $n_k f_{kk}$ *is a fixed scalar. Since* $\hat{\beta}(s)=\hat{b}(s)/(n_k f_{kk}), s\in P_k$, *the result follows.*

Several extensions of the above result are possible. The case with random (but independent) sampling over space is discussed in the online supplementary material. Here we discuss briefly the case of endogenous spatial weights. The functional principal components estimator (16) is based on orthogonality of the error ε with both *X* and *W*. This estimator is consistent only when *W* is exogenous. If *W* is endogenous, then an instrument is required. Such a functional instrument *V* has to be strictly exogenous but correlated with the functional regressor X_i^{**}. The instrument *V* may, for example, be based on a weights matrix where the elements are functions of geographic distances, which are exogenous by construction. Kelejian and Piras (2014) consider an application to demand for cigarettes in the USA, where consumers living close to the border of a state can travel some

distance into the neighbouring state to buy their tobacco. However, they would do so only if the travel distance is small and the prices in the neighbouring state are lower. This implies a weights matrix that is a combination of geographic distances and prices, and is endogenous because prices are endogenously determined. Endogenous spatial weights can also arise if the weights matrix is estimated using the same data. For example, in the context of our application here, a natural choice is the estimator of a symmetric spatial weights matrix proposed in Bhattacharjee *et al.* (2012).

A natural extension of the our estimation to the endogenous functional regressor case is based on the covariance function of *V*. Note that, in the case of simple linear regression where the OLS is given by $\hat{b}_{OLS} = Y'X / X'X$, the corresponding IV estimator is $\hat{b}_{IV} = Y'V / X'V$. As before, define

$$\hat{K}_V(s,t) = \frac{1}{n}\sum_{i=1}^{n}\{V_i(s)-\overline{V}(s)\}\{V_i(t)-\overline{V}(t)\} = \sum_{j=1}^{\infty}\hat{\kappa}_j^{(V)}\hat{\phi}_j^{(V)}(s)\hat{\phi}_j^{(V)}(t);$$

$$\hat{g}_{Y,V}(s) = \frac{1}{n}\sum_{i=1}^{n}\{Y_i-\overline{Y}\}\{V_i(s)-\overline{V}(s)\}; \text{ and}$$

$$\hat{G}_{V,X^{**}}(s,t) = \frac{1}{n}\sum_{i=1}^{n}\{X_i^{**}(s)-\overline{X}^{**}(s)\}\{V_i(t)-\overline{V}(t)\}.$$

Then, we propose the following functional principal components IV estimator of *b(.)*:

$$\hat{b}_{IV}(s) = \left[\sum_{j=1}^{m}\hat{b}_j\hat{\phi}_j(s)\right] / \left[\sum_{j=1}^{m}\hat{B}_j\hat{\phi}_j(s)\right], \tag{17}$$

where *m* is the spectral tuning parameter, $\hat{b}_j = \hat{\kappa}_j^{-1}\hat{g}_j$, $\hat{g}_j = \int \hat{g}_{Y,V}\hat{\phi}_j$, and

$$\hat{B}_j = \hat{\kappa}_j^{-1}\hat{G}_j, \quad \hat{G}_j = \int\int \hat{G}_{V,X^{**}}(s,t)\hat{\phi}_j(s)\hat{\phi}_j(t).$$

The numerator and denominator in (17) are respectively the principal components estimators (Hall and Horowitz, 2007) of the functional regression of *Y* on the functional instrument *V*, the regression of X^{**} on *V*. Then, under an instrument validity condition, the estimator in (17) is consistent for *b(s)*.

Assumption 4: ***Technical assumptions for functional IV regression inference.***

(a) *Instrument validity:* $G(s,t) = E[X^{**}(s) - EX^{**}(s)]\,[V(t) - EV(t)] \neq 0 \text{ for all } (s,t) \in S \times S.$

(b) *Technical assumptions in Hall and Horowitz (2007) hold for both the functional regressions: Y on V, and* X^{**} *on V.*

i) V has finite fourth moments. The errors in the above two regressions are identically distributed

with zero mean and finite variance not exceeding some constant C.

ii) *Consider the Karhunen-Loève expansion of the random function V:* $V - E(V) = \sum_{j=1}^{\infty} \xi_j \phi_j$, *where the* ξ_j *are pairwise uncorrelated random variables that have zero means and variances* κ_j *that are eigenvalues of the expansion. The* κ_j *satisfy the spacing condition* $\kappa_j - \kappa_{j+1} \geq C^{-1} j^{-\alpha-1}$ *for all j and some exponent α>1.*

iii) *Let* $b_j = \kappa_j^{-1} g_j$ *and* $B_j = \kappa_j^{-1} G_j$, *where* $g_j = \int g \phi_j$, $g(u) = E[(Y - EY)(V(u) - EV(u))]$ and $G_j = \int\int G(s,t)\phi_j(s)\phi_j(t)$. *The b_j's and B_j's satisfy* $\max\{|b_j|, |B_j|\} \leq Cj^{-\delta}, \delta > \frac{1}{2}\alpha + 1$.

iv) *The tuning parameter m increases with n such that* $m / n^{1/(\alpha+2\delta)}$ *is bounded away from zero and infinity.*

Assumption 4(b) is very similar to Assumption 3. Assumption 4(a) imposes non-zero covariance between the functional regressor and the instrument everywhere over the spatial domain *S*. This may be a strong assumption, but can be relaxed at the cost of analytical complexity.

Corollary 2: *Under Assumption 4,* $\hat{\beta}_{IV}(s) \xrightarrow{P} \beta(s)$ *where* $\hat{\beta}_{IV}(s) = \hat{b}_{IV}(s)/(n_k f_{kk}), s \in P_k$.

Proof: *By Theorem 1, the numerator and denominator of* $\hat{b}_{IV}(s)$ *converge to the respective functional regression coefficients, and hence the ratio converges in probability. That is,* $\hat{b}_{IV}(s) \xrightarrow{P} b(s)$. *Then the proof follows as in Corollary 1, noting that $n_k f_{kk}$ is a fixed number.*

Optimal choice of instruments is possible in this context, and weak instrument robust inference is a potential area of future research as well as alternative estimators; see the online supplementary material for further discussion.

In the remainder of this paper, including the empirical application, we focus on an exogenous weights matrix. However, the proposed framework allows for potential endogeneity of spatial structure. This extends the literature substantially. The traditional spatial econometrics literature has focussed either on spatial dependence or on spatial heterogeneity, but not both of these aspects together. The recent literature has developed methods for estimating the spatial weights, or where the spatial weights are endogenous but known *a priori*. By contrast,

this paper presents inferences for endogenous (and potentially unknown) *W* together with spatial heterogeneity.

5.2. Implementation of the Functional Principal Components Estimator

Our objective is to estimate a spatial lag model with spatial heterogeneity in the slope of a regressor, with spatial weights defined exogenously. In our empirical application, we define spatial weights by a kernel function, as in (11). Inferences are conducted by expressing the spatial lag model in reduced form as a functional regression model (6), where the functional regressor has the form given by (12).

First, the spatial domain *S* is partitioned into a large number (*K*) of small areas, denoted $\{I_1, I_2, \ldots, I_K\}$. Next, we obtain average values of the regressor (hedonic characteristic) in each of these *k* locations, combined into a spatial vector $(\bar{x}_1, \bar{x}_2, \ldots, \bar{x}_K)$. Finally, we conduct functional principal components on this vector of spatial averages. However, the above vector does not have any cross section variation. This is because the cross section variation in x_i is sacrificed in the process of aggregation by local averaging. To recover this information, we replace $\bar{x}_k$, for observation $i \in I_k$, with

$$X_k^{**} = \bar{x}_k + x_i \frac{1 - f_{0i}}{n_k f_{0i}} = \frac{1}{n_k f_{0i}} \left[x_i (1 - f_{0i}) + n_k f_{0i} \bar{x}_k \right],$$

where $f_{0i} = f_{h,i}(I_k)$ is the modal kernel density centred on the location of *i*, and n_k denotes the sample size in partition I_k. Correspondingly, we transform the response variable (*y*) into local mean deviations: $y_i^* = y_i - \bar{y}_k, \quad i \in I_k.$ Then, functional regression proceeds by obtaining a small number of functional principal components and regressing the transformed response variable (y^*) on these principal components.

The steps of the estimation method are as follows:

1. Partition the spatial domain (territory), *S*, into *k* potential submarkets, denoted $\{I_1, I_2, \ldots, I_K\}$. For each house *i*, identify the partition *j* to which it belongs: $i \in I_k$.

2. Construct the response variable as $y_i^* = y_i - \bar{y}_j, \quad i \in I_k,$ and the functional average surface as

$$X_i^{**} = \left(x_1^{**}, x_2^{**}, \ldots, x_K^{**}\right) = \left(\bar{x}_1, \bar{x}_2, \ldots, \bar{x}_k + x_i \frac{1 - f_{0i}}{n_k f_{0i}}, \ldots, \bar{x}_K \right).$$

3. Conduct functional principal components on X_i^{**}, estimate the *m* principal component factors $\left(\hat{\phi}_1, \hat{\phi}_2, \ldots, \hat{\phi}_m\right), m << K$, with corresponding eigenvalues $\hat{\kappa}_1 > \hat{\kappa}_2 > \ldots > \hat{\kappa}_m >> 0$.

4. Obtain the functional principal components estimator as (16): $\hat{b}(I_k) = \sum_{j=1}^{m} \hat{b}_j \hat{\phi}_j(I_k), \quad k = 1, \ldots, K,$

where $\hat{b}_j = \hat{\kappa}_j^{-1} \hat{g}_j$, $\hat{g}_j = \int \hat{g} \hat{\phi}_j$, and $\hat{g}(I_k) = \frac{1}{n} \sum_{i=1}^{n} \{Y_i^* - \overline{Y}^*\} \{X_i^{**}(I_k) - \overline{X}^{**}(I_k)\}$.

5. Finally obtain the estimated FPC surface of the functional regression coefficient as

$$\hat{\beta}(I_k) = \hat{b}(I_k) / (n_k f_{0i}) .$$

The proposed estimation methodology can now be applied to the data.

5.3. Submarket Delineation by Spatial Clustering

Once the functional regression slope surface is estimated, Theorem 1 suggests using the surface $\hat{\beta}(s)$ and $\overline{x}(s)$ to compute housing submarkets by spatial clustering. The notion of clustering here is related to projections on the effective dimension reduction (EDR) space (Li and Hsing, 2010). Based on the importance accorded to spatiality, there are several ways such clustering can be used: spatial clustering (Knorr-Held and Raßer, 2000); clustering based only on similarities in functional variables (Booth *et al.*, 2008); clustering based on a combination of spatial proximity and similarity in characteristic space (Zhang et al., 2014); or spatial clustering based on heterogeneous slope (Castro et al., 2015).

Thus, we propose a two-stage procedure for submarket delineation. In the first stage, we estimate the functional surfaces $\hat{\beta}(s)$ and $\overline{x}(s)$, by spatial functional regression and spatial local averaging, respectively. Then, in the second stage, we estimate submarkets by applying Ward's aggregative clustering jointly to $\hat{\beta}(s)$ and $\overline{x}(s)$. This iterative method proceeds by joining at each step the two subclusters that result in the minimum increase in the degree of within-cluster heterogeneity (sum of squares); see Everitt (1993). Theorem 1 shows that submarket delineation should be conducted by spatial clustering. However, the full development of spatial clustering methods in a spatial functional setting lies outside the domain of the current research, and is retained for future work. Nevertheless, the clusters estimated in our application are observed to have a strong spatial orientation, which relaxes the necessity to conduct spatial clustering in our case.

6. Application to the Aveiro-Ílhavo Urban Housing Market in Portugal

In this section, we apply the proposed methodology to delineate housing submarkets in a specific urban housing market – the neighbouring municipalities of Aveiro and Ílhavo in central Portugal. The municipalities of Aveiro

and Ílhavo have areas of 200 km² and 75 km² respectively, and population of 78,454 and 38,317 inhabitants respectively (Census of Portugal, 2011). Excluding the lagoon, the population density is 600 inhabitants per km², which is typical for an urban agglomeration in Portugal.

The above spatial domain is divided into the following main zones, each representing aggregation of smaller administrative areas with relatively homogeneous neighbourhoods and house prices (Figure 1): i) The inner city of Aveiro, with a population of 32,000 inhabitants – the core of the urban municipality; ii) The smaller city of Ílhavo, with a population of 5,000 inhabitants, the second urban centre of the agglomeration; iii) A semi-rural area with 30,000 inhabitants, where a significant part of the land is used for agriculture, but almost the entire population works in the manufacturing and service sectors, and housing constitutes a mixture of new urban developments with old rural settlements; iv) A suburban area with 33,500 inhabitants spread around the city of Aveiro, with a settlement and employment pattern similar to the above semi-rural area but with a higher proportion of new urban settlements; v) Gafanha da Nazaré, the port area with a population of 13,000 inhabitants, characterized by a mix of industrial and residential areas; and vi) The seaside resorts Barra and Costa Nova, with a permanent population of 3,000 inhabitants and where secondary residences and holiday rental properties dominate. As the above description shows, the Aveiro-Ílhavo urban housing market has enough variation over space to enable use of the proposed methods and framework to delineate submarkets.

The database was provided by the firm Janela Digital S.A., which owns and manages the real estate portal database Casa Sapo – the largest site in Portugal for real estate advertisement. The data pertain to the time period October 2000 and March 2010 and include around 4 million records of properties available for transaction. For the specific case of Aveiro and Ílhavo, the database included 47,188 different properties. This empirical work used 12,467 observations on completed transactions; for details on cleaning of data and omission of incomplete cases, see Bhattacharjee *et al.* (2012).

In addition to the price of each property, the database includes two main categories of variables for each dwelling: i) the intrinsic physical attributes, and ii) the location and neighbourhood of the building. The first group includes number of rooms, state of preservation (restoration), age of construction and area (living space, built area, etc.). A set of other physical housing characteristics, obtained from a free text field where real estate advertisers describe the property, was also used. The second group of attributes relates to housing location and to the characteristics of the neighbourhood, aggregated into a set of distances from different urban, local utility,

recreation and transport facilities; see Bhattacharjee *et al.* (2012) for a full discussion. Since only a small proportion of houses were fully geo-referenced, the houses were placed within into the smallest homogeneous areas that the database can describe, and the centres of the 76 such areas were geo-referenced; see online supplementary material for further detail.

The data reflect large variation in housing characteristics (Bhattacharjee *et al.*, 2012). The average price (Euros per square meter) is 1,126, and ranges from 178 up to 5,714 across the 76 zones. The average living area across the 76 zones is 149 m^2, varying between 20 m^2 and 600 m^2. 28.4% of the sample dwellings are single houses, 71.6% are flats and 12.3% are duplex (flats with two floors); 39.3% have a balcony, 18.2% have a terrace, 16.1% have garage space, and 10.3% have a garage; 43.3% have central heating while 28.9% have a fireplace. Location attributes show large spatial variation as well. On average, houses are located at 3.2 km from the CBD, while the maximum distance to the CBD is 16 km.

In order to capture the main dimensions of the housing characteristics, maximum likelihood factor analysis with orthogonal varimax rotation was applied to the hedonic housing attributes. Thus, the hedonic features were organised into 5 factors, which together explain 54% of the total variation in 43 hedonic characteristics. The factors provide clear interpretation in terms of behavioural collections of housing characteristics: of the 5 factors, 3 relate to location attributes (factor 1 - accessibility to the centre or central amenities; factor 2 - accessibility to local amenities; factor 3 - accessibility to beaches) and the other two represent the intrinsic attributes of dwellings (factor 4 - housing dimension; and factor 5 - additional desirable features).[5]

We estimate a hedonic model using the above data, modeling house prices as a function of living area, the above 5 factors, and time on the market. In this paper, specific attention is focused on living area, the shadow price for which is expected to vary over the spatial domain, and may be considered a good candidate to analyse housing spatial segmentation in the Aveiro-Ílhavo area. Therefore, our functional regression slope, $\beta(s)$, corresponds to living area, and the remaining attributes are assumed to have spatially fixed coefficients.[6]

Our central inference is reported in a plot based on clustering along two different characters that represent the spatial housing segmentation for Aveiro and Ílhavo; detailed analyses are included in the online supplementary

[5] Table 11 in Bhattacharjee *et al.* (2012) reports a detailed description of the factors.

[6] The prices reported in the dataset are asking prices and not final transaction prices. Time on the market is included to capture the wedge between asking and final prices (Bhattacharjee *et al.*, 2012).

material. For this purpose, we apply cluster analysis (Everitt, 1993) jointly to: i) living area (measured in square meters) averaged across all houses within each zone, $\bar{x}(s)$; and ii) the estimated functional regression coefficient $\hat{\beta}(s)$, representing the shadow price of living space (Figure 2), using the methods developed in section 5. Theorem 1 (section 3) emphasizes spatial clustering, which we do not explicitly apply here. However, the clusters reflect clear spatial concentration.

The spatial distribution of living area (Figure 2) shows a distinction between the smaller space in inner urban areas and the increasing available area as we move towards the periphery. The beach areas, secondary urban centres and main roads distort somewhat this regular concentric pattern. Within the inner city there is a distinction between areas with old traditional buildings and social houses (with the lowest living space) and more modern and affluent residential areas; there is a similar contrast between Barra and Costa Nova beaches. The smooth spatial variation in average living area suggests the application of functional principal components. We construct X^{**} and conduct spectral decomposition.

Next, we construct our dependent variable controlling for additional regressors and spatial fixed effects. We conduct fixed effects regression for the logarithm of price per square meter (*y*) on the 5 factors, plus time-on-the-market, allowing for zone-level fixed effects. The regressor slopes are assumed fixed, not spatially varying. The residuals constitute our modified dependent variable, y^*, for functional regression.

Based on an exogenous distance-based spatial weights using a bivariate Gaussian kernel and the spectral decomposition of the covariance function of X^{**}, we obtain our functional regression estimates, first of $\hat{b}(I_k)$, and then $\hat{\beta}(I_k), k = 1,\ldots,76$. From these estimates, we infer the estimated spatially varying living area elasticity of price. Note that, the response variable here is logarithm of price per unit living area, and not the logarithm of price in itself. Hence the estimated elasticity for zone *k* is given by $1+\hat{\beta}(I_k)$. Finally, we conduct cluster analysis on these shadow prices, and report the spatial pattern in Figure 2.[7] A concentric pattern of the shadow prices is also evident. The shadow price of living space is highest in the city centre and decrease as we move towards the periphery, meaning that the premium for a larger house is higher in the more urban areas.

[7] Zone-boundaries are demarcated by Voronoi tessellations (Okabe *et al.*, 2000), as convex polygons from the intersection of half-spaces between centres of neighbouring zones.

The above regular concentric pattern is punctuated by four exceptions: i) areas corresponding to urban expansion along the main axial roads; ii) the urban centre of Ílhavo; iii) the urban centre of Gafanha; and iv) the Barra seaside resort, where the predominant new flats have a relatively strong premium for a larger apartment. Conversely, Costa Nova, a resort to the south of Barra, has mainly traditional small houses with rigid dimensions, attracting a very low premium for extra size (demand for a nice location and style of houses, and not so much for larger living space). The dominant pattern is one of inverse relationship between $\bar{x}(s)$ and $\hat{\beta}(s)$; the lower the average living space, the higher the shadow price. As a consequence, and in line with Theorem 1, the submarkets presented in Figure 2 conform to the two delineation principles – similarity in hedonic characteristics and similarity in hedonic prices.

The resulting submarkets, ordered by decreasing value of average living space, show a concentric pattern, with some interesting features. The urban core of Aveiro corresponds to the submarket 6, with the smallest living area and the highest premium for additional space; submarket 4 corresponds to the outer ring of Aveiro, with extensions along the main roads, but also to some inner city areas (Gulbenkian and Bairro do Liceu) with relatively large high quality houses; submarket 5 corresponds to the previously discussed case of Costa Nova and three other areas with limited residential use, where the reduced living space is coupled with very low marginal returns to space; the remaining submarkets reflect the expected pattern of peripheral areas.

In summary, the submarkets obtained from the above analysis, based on clustering jointly along two dimensions, $\bar{x}(s)$ and $\hat{\beta}(s)$, produces submarkets that have a clear spatial context and approximately concentric pattern around the CBD of Aveiro. However, this concentric pattern is punctuated by processes of urban development – beach areas, secondary urban centres and main axial roads – that in turn reflect historical processes of development of the urban area.

The results are based on a new functional regression framework and methodology accounting both for spatial dependence and spatial heterogeneity. In our empirical analysis, the spatial weights matrix is defined exogenously by a geographical-distance based independent bivariate Gaussian kernel. However, one can equally use estimated spatial weights, where a natural choice may be the estimator proposed in Bhattacharjee *et al.* (2012). However, in this case, the spatial weights and functional regressor will be endogenous. The IV estimator proposed here provides very similar submarket delineation, and the results are not reported separately.

7. Conclusion

The main topic of this paper was the definition of housing submarkets in terms of its conceptualization and empirical delineation. A new framework and methods based on functional data analysis were developed, integrating ideas and approaches from functional data analysis, spatial econometrics and locally weighted regressions. This allows for spatial dependence and spatial heterogeneity, and can accommodate endogenous spatial weights. In allowing for endogenously determined submarkets and endogenous spatial regression, our work addresses important limitations of existing methods.

In the literature, analysis of housing segmentation has been conducted in several ways: i) by similarity of hedonic housing characteristics, ii) by similarity of hedonic prices, or iii) by the degree of substitutability of housing units. We show that spatial clustering based on i) and ii) also imply iii). In our application to an urban housing market in Portugal, clustering by housing characteristics and shadow prices partly overlap, and spatial clustering based on both produces submarkets where houses are substitutable.

The proposed synthesis and corresponding methods extend the literature along several directions. First, the framework can allow spatial structure and submarkets to evolve endogenously. Second, our framework extends FDA tools and methods to the spatial domain, and specifically the spatial lag model with spatial heterogeneity in slopes and spatial fixed effects. Third, once such submarkets have been delineated, spatial dependence can be examined by estimating cross- and within-submarket spatial weights (Bhattacharjee *et al.*, 2012).

Several further research problems develop from our work. First, while our framework aids analyses of endogenously produced submarkets, finding the asymptotic convergence rates for the proposed IV estimator for the functional regression model is retained for future work. Inferences robust to weak instruments in this setting may also be useful. Further, relaxing the fixed design assumption will enhance applicability of the methods.

Second, combining the proposed approach with estimated spatial weights is a topic for further research. For conducting inference on W, one can exploit the fact that the error term in the reduced form is $(I - W)^{-1}\varepsilon$, and hence the spatial autocovariance matrix of these errors is a 1–1 function of a symmetric weights matrix (Bhattacharjee and Jensen-Butler, 2013). Further, the submarkets identified in the previous step can also be used to estimate within and between submarket spatial weights (Bhattacharjee *et al.*, 2012).

Finally, estimation and inferences on an unknown spatial weights matrix W has another important advantage. FDA on the partial effects can then borrow strength over the network (defined by W) using ideas and concepts

from small area statistics; see, for example, Castro *et al.* (2015). Perhaps most importantly, the proposed framework offers the possibility of studying the endogenous evolution of urban spatial structure. All these lines of future research are exciting.

References

Adair, A., Berry, J. and McGreal, W. (1996). Hedonic modelling, housing submarkets and residential valuation. *Journal of Property Research* **13**(1), 67-83.

Anas, A., Arnott, R. and Small, K. (1998). Urban spatial structure. *Journal of Economic Literature* **36**(3), 1426-1464.

Anselin, L. (1988). *Spatial Econometrics: Methods and Models*. Kluwer: Boston.

Anselin, L. and Lozano-Gracia, N. (2008). Errors in variables and spatial effects in hedonic house price models of ambient air quality. *Empirical Economics* **34**, 5-34.

Anselin, L., Lozano-Gracia, N., Deichmann, U. and Lall, S. (2010). Valuing access to water: a spatial hedonic approach, with an application to Bangalore, India. *Spatial Economic Analysis* **5**(2), 161-179.

Bailey, N., Holly, S. and Pesaran, M.H. (2014). Modelling Spatial Dependence with Pairwise Correlations. *Journal of Applied Econometrics* (Forthcoming).

Bhattacharjee, A., Castro, E.A. and Marques, J.L. (2012). Spatial interactions in hedonic pricing models: the urban housing market of Aveiro, Portugal. *Spatial Economic Analysis*, **7**(1), 133-167.

Bhattacharjee, A. and Holly, S. (2013). Understanding interactions in social networks and committees. *Spatial Economic Analysis* **8**(1), 23-53.

Bhattacharjee, A. and Jensen-Butler, C. (2013). Estimation of the spatial weights matrix under structural constraints. *Regional Science and Urban Economics* **43**, 617-634.

Booth, J.G., Casella, G. and Hobert, J.P. (2008). Clustering using objective functions and stochastic search. *Journal of the Royal Statistical Society B* **70**, 119-140.

Bourassa, S.C., Hamelink, F., Hoesli, M. and MacGregor, B.D. (1999). Defining housing submarkets. *Journal of Housing Economics* **8**, 160-183.

Bourassa, S.C., Hoesli, M. and Peng, V.C. (2003). Do housing submarkets really matter?. *Journal of Housing Economics* **12**(1), 12-28.

Cai, T. and Hall, P. (2006). Prediction in functional linear regression. *Annals of Statistics* **34**(5), 2159-2179.

Castro, E.A., Zhang, Z., Bhattacharjee, A., Martins, J.M. and Maiti, T. (2015). Regional Fertility Data Analysis: A

Small Area Bayesian Approach. In: D.K. Dey, A. Loganathan, U. Singh and and S.K. Upadhyay (Eds.), *Current Trends in Bayesian Methodology with Applications*, Chapman & Hall/CRC Press (In Press), Chapter 10.

Census of Portugal (2011). Recenseamento da população e habitação, Instituto nacional de estatística (http://www.ine.pt/).

Dale-Johnson, D. (1982). An alternative approach to housing market segmentation using hedonic price data. *Journal of Urban Economics* **11**(3), 311-332.

Everitt, B. S. (1993). *Cluster Analysis*. 3rd edition. Arnold: London.

Fotheringham, A., Brunsdon C. and Charlton, M. (1998). Geographically weighted regression: a natural evolution of the expansion method for spatial data analysis. *Environment and Planning A* **30**, 1905-1927.

Fujita, M. and Thisse, J. (2002). *Economic of Agglomeration: Cities, Industrial Location, and Regional Growth*. Cambridge University Press.

Galster, G. (2001). On the nature of neighbourhood. *Urban Studies* **38**(12): 2111-2124.

Gillen, K., Thibodeau, T. and Wachter, S. (2001). Anisotropic autocorrelation in house prices. *Journal of Real Estate Finance and Economics* **23**(1), 5-30.

Goodman, A. C. and Thibodeau, T. G. (2007). The spatial proximity of metropolitan area housing submarkets. *Real Estate Economics* **35**(2), 209-232.

Grigsby, W., Baratz, M., Galster, G., and Maclennan, D. (1987). The dynamics of neighborhood change and decline. *Progress in Planning* **28**(1), 1–76.

Guillas, S. and Lai, M.J. (2010). Bivariate splines for spatial functional regression models. *Journal of Nonparametric Statistics* **22**, 477-497.

Hall, P. and Horowitz, J.L. (2007). Methodology and convergence rates for functional linear regression. *Annals of Statistics* **35**(1), 70-91.

Kelejian, H.H. and Piras, G. (2014). Estimation of spatial models with endogenous weighting matrices, and an application to a demand model for cigarettes. *Regional Science and Urban Economics* **46**, 140-149.

Knorr-Held, L. and G. Raßer (2000). Bayesian detection of clusters and discontinuities in disease maps. *Biometrics* **56**(1), 13-21.

Lancaster, K.J. (1966). A new approach to consumer theory. *Journal of Political Economy* **74**(2), 132-157.

Lefebvre, H. (1974 [1991]). *The Production of Space*. (Trans.) Nicholson-Smith, D., Blackwell: Oxford.

LeSage, J. and Pace, R. (2009). *Introduction to Spatial Econometrics*. Chapman & Hall/CRC.

Li, Y. and Hsing, T. (2010). Deciding the dimension of effective dimension reduction space for functional and high-dimensional data. *Annals of Statistics* **38**(5), 3028-3062.

Malpezzi, S. (2003). Hedonic pricing models: a selective and applied review. Chapter 5, In: Gibb, K. and O'Sullivan, A. (Eds.), *Housing Economics and Public Policy: Essays in Honour of Duncan Maclennan*, Blackwell Science: Oxford (UK), 67-89.

McMillen, D.P. (1996). One hundred and fifty years of land values in Chicago: a nonparametric approach. *Journal of Urban Economics* **40**, 100-124.

Okabe, A., Boots, B., Sugihara, K. and Chiu, S.N. (2000). *Spatial Tessellations: Concepts and Applications of Voronoi Diagrams*. 2nd edition. John Wiley.

Pesaran, M.H. (2006). Estimation and inference in large heterogenous panels with multifactor error structure. *Econometrica* **74**, 967-1012.

Pesaran, M.H. and Tosetti, E. (2011). Large panels with common factors and spatial correlation. *Journal of Econometrics* **161**, 182-202.

Pryce, G. (2013). Housing submarkets and the lattice of substitution. *Urban Studies* **50**, 2682-2699.

Ramsay, J.O. and Silverman, B.W. (2005). *Applied Functional Data Analysis*. Springer-Verlag: New York.

Rosen, S. (1974). Hedonic prices and implicit markets: product differentiation in pure competition. *Journal of Political Economy* **82**, 34-55.

Rothenberg, J., Galster, G., Butler, R.V. and Pitkin, J.K. (1991). *The Maze of Urban Housing Markets: Theory, Evidence and Policy*. University of Chicago Press.

Watkins, C. (2001). The definition and identification of housing submarkets. *Environment and Planning A* **33**(12), 2235-2253.

Zhang, Z., Lim, C.-Y. and Maiti, T. (2014). Analyzing 2000-2010 Childhood Age-Adjusted Cancer Rates in Florida: A Spatial Clustering Approach. *Statistics and Public Policy* **1**(1), 120-128.

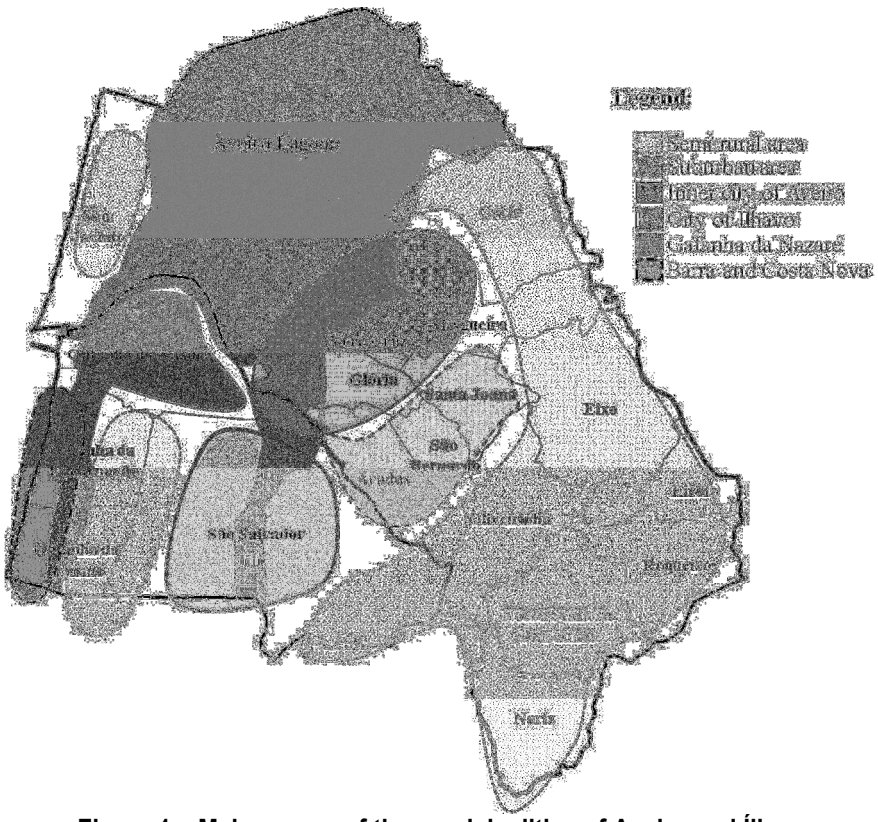
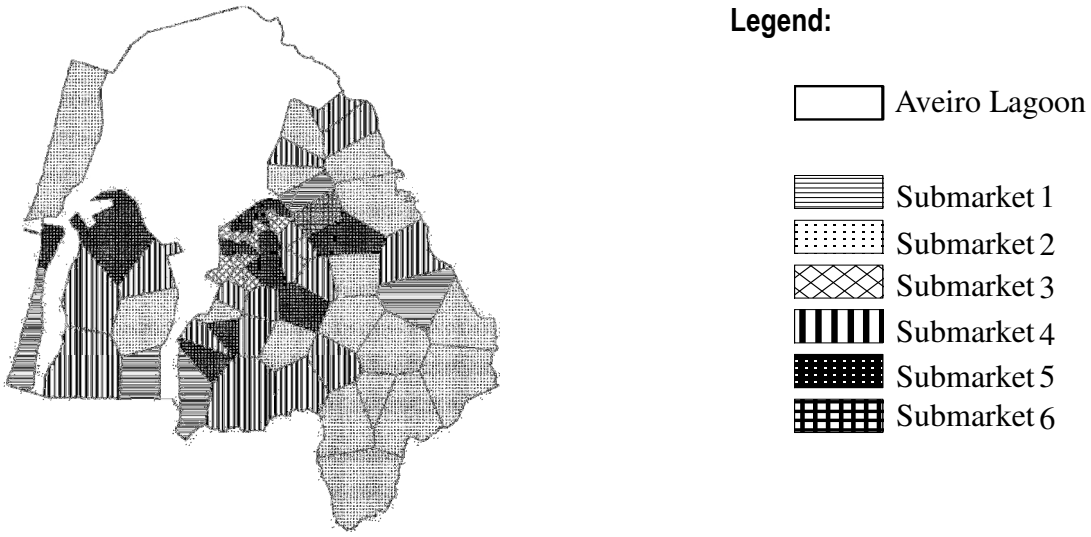


Figure 1 – Major zones of the municipalities of Aveiro and Ílhavo



	Number of zones	FDA elasticity (standardized values)	Ln Area m² (standardized values)
Submarket 1	2	-0.156	2.839
Submarket 2	17	-1.050	1.124
Submarket 3	18	-0.268	0.402
Submarket 4	20	0.782	-0.524
Submarket 5	5	-1.137	-0.948
Submarket 6	14	0.931	-1.201

Figure 2 - Submarkets based on Ward's linkage clusters: Housing characteristics combined with marginal returns to living space